Can Artificial Intelligence Pass the American Board of Orthopaedic Surgery Examination? Orthopaedic Residents Versus ChatGPT

Year: 2023 | Citations: 103 | Authors: Zachary C Lum

Abstract

Abstract Background Advances in neural networks, deep learning, and artificial intelligence (AI) have progressed recently. Previous deep learning AI has been structured around domain-specific areas that are trained on dataset-specific areas of interest that yield high accuracy and precision. A new AI model using large language models (LLM) and nonspecific domain areas, ChatGPT (OpenAI), has gained attention. Although AI has demonstrated proficiency in managing vast amounts of data, implementation of that knowledge remains a challenge. Questions/purposes (1) What percentage of Orthopaedic In-Training Examination questions can a generative, pretrained transformer chatbot (ChatGPT) answer correctly? (2) How does that percentage compare with results achieved by orthopaedic residents of different levels, and if scoring lower than the 10th percentile relative to 5th-year residents is likely to correspond to a failing American Board of Orthopaedic Surgery score, is this LLM likely to pass the orthopaedic surgery written boards? (3) Does increasing question taxonomy affect the LLM's ability to select the correct answer choices? Methods This study randomly selected 400 of 3840 publicly available questions based on the Orthopaedic In-Training Examination and compared the mean score with that of residents who took the test over a 5-year period. Questions with figures, diagrams, or charts were excluded, including five questions the LLM could not provide an answer for, resulting in 207 questions administered with raw score recorded. The LLM's answer results were compared with the Orthopaedic In-Training Examination ranking of orthopaedic surgery residents. Based on the findings of an earlier study, a pass-fail cutoff was set at the 10th percentile. Questions answered were then categorized based on the Buckwalter taxonomy of recall, which deals with increasingly complex levels of interpretation and application of knowledge; comparison was made of the LLM's performance across taxonomic levels and was analyzed using a chi-square test. Results ChatGPT selected the correct answer 47% (97 of 207) of the time, and 53% (110 of 207) of the time it answered incorrectly. Based on prior Orthopaedic In-Training Examination testing, the LLM scored in the 40th percentile for postgraduate year (PGY) 1s, the eighth percentile for PGY2s, and the first percentile for PGY3s, PGY4s, and PGY5s; based on the latter finding (and using a predefined cutoff of the 10th percentile of PGY5s as the threshold for a passing score), it seems unlikely that the LLM would pass the written board examination. The LLM's performance decreased as question taxonomy level increased (it answered 54% [54 of 101] of Tax 1 questions correctly, 51% [18 of 35] of Tax 2 questions correctly, and 34% [24 of 71] of Tax 3 questions correctly; p = 0.034). Conclusion Although this general-domain LLM has a low likelihood of passing the orthopaedic surgery board examination, testing performance and knowledge are comparable to that of a first-year orthopaedic surgery resident. The LLM's ability to provide accurate answers declines with increasing question taxonomy and complexity, indicating a deficiency in implementing knowledge. Clinical Relevance Current AI appears to perform better at knowledge and interpretation-based inquires, and based on this study and other areas of opportunity, it may become an additional tool for orthopaedic learning and education.